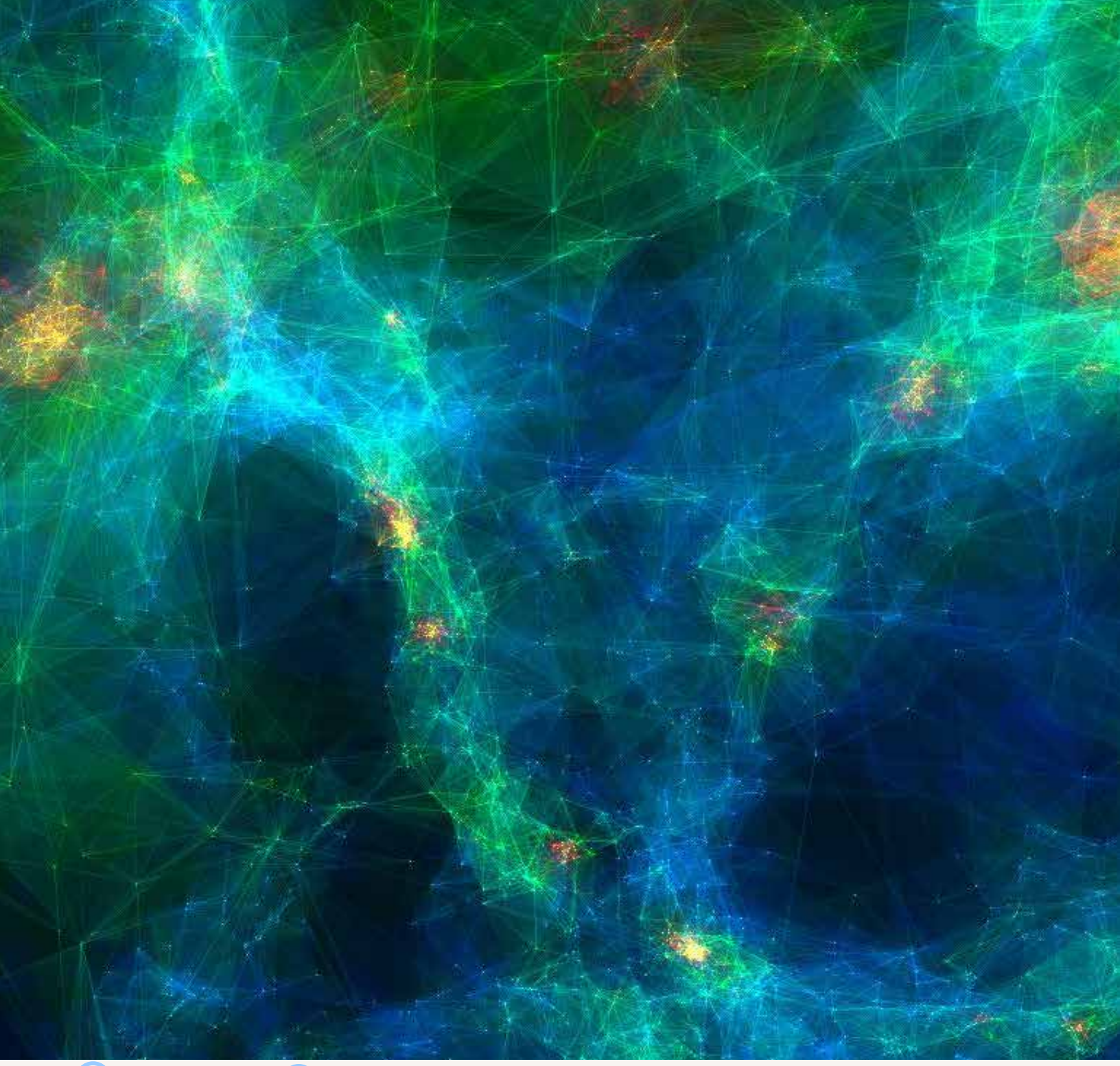


idies

Oct. 1, 2013 - Sept. 30,

Annual Review 2014

JOHNS HOPKINS UNIVERSITY | Institute for Data Intensive Engineering and Science

MESSAGE FROM THE

DIRECTOR

Recognizing the strategic importance of computing and data across the whole university, President Ronald Daniels decided a year ago to substantially broaden the scope of IDIES. Now the Institute includes five schools: the Krieger School of Arts and Sciences, the Whiting School of Engineering, the Sheridan Libraries, the School of Medicine and the Bloomberg School of Public Health.

Besides serving in a leadership role for Big Data initiatives for the University, IDIES has become responsible for the research computing efforts at JHU. In March 2015 we will open the High Performance Research Computing Facility (HPRCF). The HPRCF will be a new, world class research computing facility on the Johns Hopkins Bayview campus, partnering with the University of Maryland College Park.

Today IDIES involves over 84 faculty and more than a hundred graduate students. Over the last 12 months we have awarded 9 seed grants in a broad spectrum of topics, connecting researchers from different fields who share an interest in the science of Big Data. The past year has seen IDIES develop into a major interdisciplinary program, a large, diverse effort where faculty and students work together to solve amazing data-intensive problems, from genes to galaxies.

TABLE OF CONTENTS

Cover Art: Delaunay-based volume rendering of an N-body simulation is used to study formation of galaxy filaments and walls. Delaunay rendering triangulates a set of points so that no point in the set is inside any triangle. It interpolates space to produce a real-time pseudo rendering of the simulation volume. The program assigns color and transparency to individual triangles as a function of the density of their three vertices. *Image by Miguel Aragon-Calvo, Johns Hopkins University.*

AGENDA

2014 Institute for Data Intensive Engineering and Science (IDIES)
Annual Symposium
October 17, 2014 8:00 – 5:00
Mudd Hall Auditorium, Room 26

8:00am Continental Breakfast – Check In

9:00am "Data Intensive Science At JHU: The First Year Of IDIES"
S. Alexander Szalay, PhD, Director IDIES, Professor Astrophysics & Computer Science, Johns Hopkins University

9:20am "The New High Performance Research Computing Facility"
Jaime E. Combariza, PhD, Director Bayview HPRCF, Associate Research Scientist, Chemistry, Johns Hopkins University

9:40am "Highlights In Big Data From the School of Public Health"
Roger Peng, PhD, Associate Professor Biostatistics, Bloomberg School of Public Health, Johns Hopkins University

9:55am *Break*

10:15am "Hopkins In Health; Fostering The Intelligent Use Of information To Advance Health"
Scott Zeger, PhD, Professor of Biostatistics, Bloomberg School of Public Health

10:30am KEYNOTE ADDRESS – **"Adventures in Little Data"**
Paul Ginsparg, PhD, Professor of Physics and Information Science, Cornell University

11:15am "Accessible, Transparent, And Reproducible Genomics With Galaxy"
James Taylor, PhD, Ralph S. O'Connor Associate Professor of Biology & Computer Science at Johns Hopkins University

11:30am "Highlights In Big Data From Sheridan Libraries"
G. Sayeed Choudhury, Associate Dean for Research Data Management, Johns Hopkins University

11:45am "The World's Largest Data Science Educational Effort: The Johns Hopkins Data Science Specialization"
Jeffrey T. Leek, PhD, Associate Professor, Biostatistics and Oncology, Johns Hopkins Bloomberg School of Public Health

12:00pm *Lunch*

1:00pm "Highlights From Data Intensive Research At WSE And First IDIES Seed Funding Program"
Charles Meneveau, PhD, Professor, Mechanical Engineering, Whiting School of Engineering, Johns Hopkins University

1:15pm The IDIES Annual Seed Funding Program

- **Yanif Ahmad, PhD**, Assistant Professor, Computer Science, Johns Hopkins University
  "SIRENIC: Stream Infrastructure for the Real-time Analysis of Intensive Care Unit Sensor Data"
- **Sarah J. Wheelan, MD, PhD** , Assistant Professor, Oncology Bioinformatics, Johns Hopkins University
  "Approaching Genomics Data from Hundreds of Dimensions Simultaneously (This is not a Faster Horse)"
- **Tamer Zaki, PhD**, Associate Professor, Mechanical Engineering, Johns Hopkins University
  "The Elusive Onset of Turbulence And The Laminar-Turbulence Interface"
- **Ben Langmead, PhD**, Assistant Professor, Computer Science, Johns Hopkins University
  "Highly Scalable Software for Analyzing Large Collections of RNA Sequencing Data"
- **Nitin Daphalapurkar, PhD** , Assistant Research Professor, Mechanical Engineering, Johns Hopkins University
  "FragData—High-fidelity Data on Dynamic Fragmentation of Brittle Materials"

2:30pm *Break*

2:50pm "Big Data: Opportunities And Challenges In Health Care"
Patricia M. Davidson, PhD, MEd, RN, FAAN, Dean, Johns Hopkins School of Nursing

3:05pm "Highlights In Big Data From SOM"
Steven L. Salzberg, PhD, Professor, Departments of Medicine & Biostatistics, Johns Hopkins University

3:20pm KEYNOTE ADDRESS – "What is the Big Data Problem in Biology?"
David J Lipman, MD, Director, National Center for Biotechnology Information (NCBI),
National Institutes of Health (NIH)

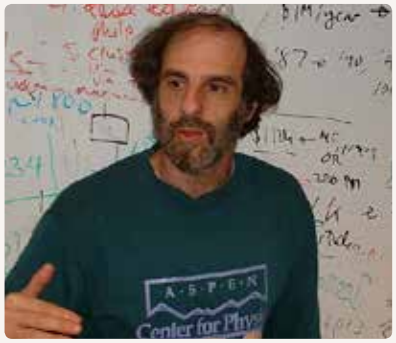
4:05pm Closing Remarks
S. Alexander Szalay, PhD, Director IDIES, Professor Astrophysics & Computer Science, Johns Hopkins University

4:10pm **Poster Session and Cocktail Hour**
Mudd Hall Commons – Upper Level

KEYNOTE SPEAKERS

PAUL GINSPARG, PhD

Professor of Physics and Information Science, Cornell University



Paul Ginsparg is a Professor of Physics and Information Science at Cornell University. He has authored papers in quantum field theory, string theory, and information science. While visiting Aspen in the summer of 1991, he started the e-print archives (now arXiv.org). He was recently named a "White House Champion of Change" for work in open access publication, and a Simons Fellow for work in theoretical physics.

DAVID LIPMAN, MD

Director, National Center for Biotechnology Information, NIH

Dr. David Lipman is the Director of the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine within the National Institutes of Health (NIH). Dr. Lipman was appointed as NCBI's first Director in 1989 and has overseen its growth into one of the most heavily used resources in the world for the search and retrieval of biomedical information, with over three million users each day. He is one of the developers of the original BLAST (Basic Local Alignment Search Tool) algorithm for rapidly identifying biological sequences that are similar to a queried sequence. Dr. Lipman is the recipient of numerous awards and is an elected member of the National Academy of Sciences, the Institute of Medicine, and the American Academy of Arts and Sciences.

SPEAKER BIOS

Jaime E. Combariza, PhD

Associate Research Scientist at the Department of Chemistry and Director of the Bayview High Performance Research Computing Facility (HPRCF).

Jeffrey T. Leek, PhD

Associate Professor Biostatistics and Oncology, Johns Hopkins Bloomberg School of Public Health. Studies statistical processing of sequencing data and using genomic data to create robust biomarkers.

Scott Zeger, PhD

Professor of Biostatistics at the Bloomberg School of Public Health. Works on statistical models of medical and public health problems, and predictive Bayesian network models of disease etiology.

James Taylor, PhD

Ralph S. O'Connor Associate Professor of Biology and Computer Science at Johns Hopkins University. Formerly Associate Professor of Biology and Mathematics & Computer Science at Emory University.

Sayeed Choudhury

Associate Dean for Research Data Management and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries and member of the Executive Committee IDIES.

SPEAKER BIOS

Roger Peng, PhD

Associate Professor in the Department of Biostatistics Johns Hopkins Bloomberg School of Public Health and is Co-Editor of the Simply Statistics Blog.

Charles Meneveau, PhD

Louis M. Sardella Professor of Mechanical Engineering, and Associate Director of IDIES. His research is on various fundamental and applied aspects of fluid turbulence.

Patricia M. Davidson, PhD, MEd, RN, FAAN

Professor and Dean of Johns Hopkins University School of Nursing. Works on developing innovative models of transitional care and improving the cardiovascular health of underserved populations.

Steven L. Salzberg, PhD

Professor of Biomedical Engineering, Computer Science, and Biostatistics, Director of the Center for Computational Biology. He develops new computational methods to analyze DNA.

Alex Szalay, PhD

IDIES Director, Alumni Centennial Professor of Astronomy, and Computer Science Department Professor. He is a cosmologist, working on Big Data.

SPEAKER BIOS

Yanif Ahmad, PhD

Assistant Professor of Computer Science. Yanif's Data Management Systems Lab aims to democratize scalable, domain-specific data systems construction via declarative programming research.

Sarah J. Wheelan, MD, PhD

Assistant Professor, Department of Oncology. Works on DNA sequence/structure/function relationships. Her interdisciplinary group crochets hyperbolic planes for a coral reef installation.

Tamer Zaki, PhD

Associate Professor in the Department of Mechanical Engineering. His research focuses on transitional flows and laminar-turbulence interfaces.

Benjamin Langmead, PhD

Assistant Professor of Computer Science. His group seeks to make high-throughput biological datasets easy for biomedical researchers to use.

Nitin P. Daphalapurkar, PhD

Assistant Research Professor with the Dept. of Mechanical Engineering. Uses material point method computations to model and control failure in materials.

UPDATES FROM AFFILIATES

HONORS & AWARDS

Data Conservancy receives Alfred P. Sloan Foundation grant to connect publications to data

The Data Conservancy, a data infrastructure development program associated with IDIES, has received an Alfred P. Sloan Foundation award to create a set of services to connect research publications with their underlying data. The new award funds a partnership to develop data curation infrastructure that will build, store, update, and retrieve connections among publications and data.

In the past, when datasets were small and publication was only in print, it was easy to publish all parts of a research project in one place, simply by including data tables in the final printed publication. Today's research publications consist of many distinct building blocks such as text, graphics, and data, which often reside in different repositories hosted at different institutions using different technologies. In this new world, preserving research literature requires not only preserving all parts of the whole, but also keeping the relationships among the parts functional. The new project will allow authors and readers to maintain these complex linkages into the indefinite future.

"We believe that the models developed as a result of this project will enable new forms of scholarly communication, and thus help to set the stage for the future of research and digital publishing," said **Sayeed Choudhury**, Associate Dean for research data management and Hodson Director of the Digital Research and Curation Center at JHU. "Our partnership represents broad perspectives and multifaceted experience, which we believe will result in more meaningful solutions that can be generalized for the entire community."

The two-year, $600,000 grant was awarded by the Alfred P. Sloan Foundation to a partnership between the Data Conservancy, IEEE and the Portico community-supported digital archives. For more information about the award, see the Portico press release linked from the IDIES web site at www.idies.jhu.edu/news/data-conservancy.

Three IDIES affiliates named highly cited researchers in 2014

Three researchers from IDIES have been named "Highly Cited Researchers" by the scientific publisher Thomson Reuters, an honor awarded to only the most influential researchers in a wide variety of scientific fields. The three researchers are **Steven Salzberg**, **Alex Szalay**, and **Ani Thakar**.

Salzberg, Szalay, and Thakar were three of just 3,215 scientists worldwide named to the prestigious list, including just 29 from Johns Hopkins institutions. Salzberg, the Director of JHU's Center for Computational Biology and Associate Director of IDIES, received the rare honor of appearing on the list twice, being named a Highly Cited Researcher in both microbiology and computer science. Szalay, the Director of IDIES, appeared on the list for the second time, having been on the prior list in 2001, as was Salzberg. Thakar, the IDIES Associate Director of Operations, is a first-time honoree. Both Szalay and Thakar were recognized in the area of space science. All three have worked together extensively on techniques and applications of big data for scientific research.

The list of Highly Cited Researchers is based on Web of Science, an online catalog of scientific publications and the citations among them. Analysts at Thomson Reuters started from a list of all papers published in each of 21 scientific fields between 2002 and 2012. They took all the authors whose names appeared on those papers and sorted them by number of citations. An author was included on the Highly Cited Researchers list only if the total citations to their listed papers during that year put them in the top 1% of all researchers in their field.

UPDATES FROM AFFILIATES

RESEARCH

New IDIES Database Offers New Opportunities for Clinical Research

IDIES affiliates **Steven Jones** and **Seth Martin** have developed the Very Large Database of Lipids (VLDL), a new resource bringing previously-inaccessible laboratory data to researchers worldwide. The innovative database has already been used to test the accuracy of a clinical rule of thumb, resulting in greatly increased diagnostic power for the highest-risk patients.

VLDL data comes from laboratory test results at Atherotech, Inc., an Alabama-based private industry diagnostic lab. The database contains de-identified laboratory measurements such as cholesterol levels of all major lipoproteins and their density subclasses, as well as other biomarkers, from 1.3 million Americans. A dataset of this size and scope is not feasible to develop using conventional methods of grant-funded research, and so the VLDL offers a powerful complement to traditional methods of medical and public health research.

VLDL data have already led to several high-impact peer-reviewed studies. One such study tested the "Friedewald Method," a common rule of thumb to help doctors estimate patients' levels of low-density lipoprotein cholesterol (LDL-C). Comparing Friedewald estimates to actual laboratory measurements of LDL-C allowed researchers to develop a more statistically-robust method to estimate this important clinical parameter. This work was published in one of the world's leading medical journals, JAMA.

Jones and colleagues are currently working with Atherotech to create a larger database called VLDL 2.0, with de-identified laboratory results from 4.4 million people. The pioneering database will include matches with mortality data from the Centers for Disease Control and Prevention (CDC), offering another opportunity to connect big data with traditional public health research.

The VLDL project is registered on ClinicalTrials.gov and invites interested investigators to contact the VLDL lead investigators Steven Jones, sjones64@jhmi.edu or Seth Martin, smart100@jhmi.edu concerning opportunities for collaboration or proposals for original research.

Tapping into Big Data

The secrets of hundreds of millions of galaxies and stars are stored in a humming, whirring computer-filled room on the first floor of the School of Arts and Sciences' Bloomberg Center for Physics and Astronomy. And they have lots of company, such as the genetic coding of loblolly pine trees (six times longer than human genetic sequences), sensor-collected soil data, and multi-terabyte data sets used to chart air turbulence in three dimensions.

"What we have here is probably hundreds of times the amount of information in the Library of Congress," says **Alex Szalay**, director of the Institute for Data Intensive Engineering and Science (IDIES), standing amid 16 racks of neatly stacked processors and disks holding a combined 10 petabytes of storage.

Through IDIES, researchers will be able to piggyback on previous efforts to collect and analyze vast data sets that combine information in entirely new ways. For example, **Steve Salzberg**, director of the School of Medicine's Center for Computational Biology, is deciphering the genome of the loblolly pine tree, which has about 22 billion base pairs.

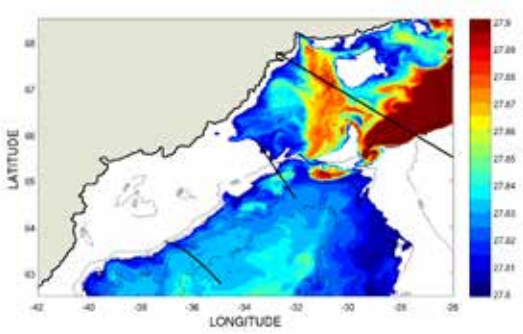
Ambitious research projects are quickly filling up the space in that computer room. That's why the university is preparing for the next stage: a High Performance Research Computing Facility slated to open in September, which will be located at Johns Hopkins Bayview Medical Center in East Baltimore.

Yet even as storage capacity increases and programs calculate information at ever-quickening speeds, the demand for more and faster seems infinite. "We won't have enough technology to handle it all," asserts Professor Szalay. "That's why we have to keep being innovative."

This article was adapted from a longer version that appeared in the Spring 2014 issue of Johns Hopkins Arts and Sciences magazine, written by Karen Nitkin.

UPDATES FROM AFFILIATES

ANNOUNCEMENTS



Fates and Travel Times of Denmark Strait Overflow Water in the Irminger Basin

Thomas Haine, Morton K. Blaustein Chair and Professor of Earth and Planetary Sciences, Assistant Research Scientist **Inga Koszalka**, and Visiting Scholar **Marcello Magaldi** have been published in the Journal of Physical Oceanography on the Fates and Travel Times of Denmark Strait Overflow Water in the Irminger Basin (Koszalka et al., 2013). The paper uses very-high resolution ocean circulation model solutions produced by the Homewood High-Performance Cluster and simulates the trajectories of more than 10,000 Lagrangian particles using the DataScope analysis cluster.

The Denmark Strait Overflow (DSO) supplies about one-third of the North Atlantic Deep Water and is critical to global thermohaline circulation in the ocean. Knowledge of the pathways of DSO through the Irminger Basin (southwest of Iceland) and its transformation there is still incomplete, however. Koszalka, Haine, and Magaldi deployed over 10,000 synthetic Lagrangian particles at the Denmark Strait (between Iceland and Greenland) in a high-resolution ocean model to study these issues. The particle trajectories show that the mean position and potential density of dense waters cascading over the Denmark Strait sill evolve consistently with hydrographic observations. The dense water pathways on the continental shelf are also consistent with observations. Particles released on the shelf in Denmark Strait constitute a significant fraction (~25%) of the dense water particles recorded at the downstream Angmagssalik monitoring section within 60 days. Some particles circulate on the shelf for several weeks before they spill off the shelf break and join the overflow from the sill. Mixing is not uniformly strong. Instead, there are two places where the DSO water density decreases rapidly due to intense turbulence: to the southwest of the sill and southwest of the Kangerdlugssuaq Trough (on the Greenland continental slope). After transformation in these places, the overflow particles exhibit a wide range of densities.

For more details visit:

- The Project Website - http://blaustein.eps.jhu.edu/~koszalka/dsow_webpage/DSOW.html
- The Published Paper - http://journals.ametsoc.org/doi/abs/10.1175/JPO-D-13-023.1
- Movie of Lagrangian Particles - http://orchard1.pha.jhu.edu/Media/Default/Research/Animation4.m4v

Affiliate Tamás Budavári joins the faculty of Applied Math and Statistics

July 1, 2014. The Department of Applied Mathematics and Statistics (AMS) of the Whiting School of Engineering is pleased to announce the addition of a distinguished new faculty member: IDIES affiliate Professor **Tamás Budavári**. After graduating with a degree in Physics from Eötvös Loránd University, Budapest, Budavári came to Baltimore to work on large astronomical experiments and astronomical Big Data. His scientific interests span a wide range of topics from survey astronomy to computational statistics. His new appointment will open up additional possibilities for exciting projects and collaborations.

Budavári comes to AMS from the Department of Physics and Astronomy in JHU's Krieger School of Arts and Sciences, where he has been a researcher since 2001. He is a key member of the Sloan Digital Sky Survey (SDSS) collaboration and the International Virtual Observatory Alliance (IVOA). Much of his recent research concerns the theory of "cross-matching" astronomical catalogs: "If you have separate collections of sources in a number of surveys, how do you know which objects observed in one correspond to which detections in the others?" He works on novel image processing strategies to improve measurements from repeated exposures by eliminating the varying blur in each snapshot, leading to a new research effort in data exploration bridging the gap between human and machine learning.

Budavári will continue to work on key astronomy projects, such as the Hubble Source Catalog, but will be seen more frequently in Whitehead Hall. For more information on Prof. Budavári's research or to discuss opportunities for collaboration please feel free to contact him at budavari@jhu.edu.

OUR MISSION

We foster education and research in the development and application of data intensive technologies to problems of national interest in physical and biological sciences and engineering. The institute provides faculty, researchers and students with the structure and resources needed to accomplish these goals.

Leadership

Intellectual leadership in addressing research challenges related to the "Science of Big Data," establishing a group that leads the world in new discoveries enabled by next-generation data sets and analytics. Provide coordination of integrative activities, such as seminar series, visitors, and so on.

Vision

Continue to provide vision and oversight to high performance and data intensive computing across all of JHU, in the spirit that has proven to be highly successful over the last four years (HHPC 1 and 2, GPU). Having a large shared facility enables leveraging needed for seeking further funding opportunities.

Growth

Given the emerging need of data analytics skills for the workforce of the future, IDIES will work with the departments to establish new masters, graduate, and undergraduate programs, minors, etc., that emphasize these new skills.

Development

Continue to develop mutually beneficial corporate partnerships and through these affiliations transform research into sustainable, real-world applications.

Incubator

An incubator for creating/curating/publishing new data sets at JHU that could be preserved within the JHU Data Archive. This would give the group an "unfair advantage," name recognition, and additional leverage, while also motivating and focusing research around challenges and and opportunities of dealing with Big Data.

Management

Management of a significant high-performance computing facility. IDIES already has some facilities, and these need to grow as the new institute applies for new funding. A large HPC resource will be a magnet attracting new JHU researchers to the institute.

IDIES is always accepting affiliates who are Faculty and Research Scientists within the Johns Hopkins community. Visit idies.jhu.edu/join for more information, and to join today!

SEED FUNDING AWARDEES

The IDIES Seed Funding Program RFP was issued for competitive awards of $25,000. The goal of the Seed Funding initiative is to provide funding for data-intensive computing projects that (a) will involve areas relevant to IDIES and JHU institutional research priorities; (b) are multidisciplinary; and (c) build ideas and teams with good prospects for successful proposals to attract external research support by leveraging IDIES intellectual and physical infrastructure.

The following projects were selected for the inaugural Seed Funding Awards in Spring, 2014:

"SIRENIC: Stream Infrastructure for the Real-time Analysis of Intensive Care Unit Sensor Data"
by Yanif Ahmad, (Dept. of Computer Science), Raimond Winslow (Dept. Biomedical Engineering), and Yair Amir, (Dept. of Computer Science)

"Alignment to The Cancer Genome Atlas Project Raw Sequencing Reads (8948 Samples and Counting)"
by Sarah Wheelan, (Dept. of Oncology) and Srinivasan Yegnasubramanian, (Dept. of Oncology)

"The Elusive Onset of Turbulence And The Laminar-Turbulence Interface"
by Tamer A. Zaki (Dept. of Mechanical Engineering) and Gregory Eyink (Applied Math and Statistics)

"Highly Scalable Software for Analyzing Large Collections of RNA Sequencing Data"
by Ben Langmead, PhD (Dept. of Computer Science) and Jeffrey Leek, PhD (Dept. of Biostatistics)

"FragData—High-fidelity Data on Dynamic Fragmentation of Brittle Materials"
by Nitin Daphalapurkar (Dept. of Mechanical Engineering), and Lori Graham-Brady (Dept. of Civil Engineering)

YANIF AHMAD

We are designing **Sirenic** as open-source data streaming infrastructure for the real-time analysis of patient physiological data in intensive care units. Sirenic exploits systems specialization and scaling capabilities enabled by our K3 declarative systems compilation framework to realize orders of magnitude data throughput gains over current generation stream and database systems. Our proposal aims at delivering a proof-of-concept data collection and analysis pipeline to support exploratory research activities in ICU healthcare, with the explicit capability to operate on live data and to empower alarms research and event detection in the real-time setting.

SARAH WHEELAN

Alignment to The Cancer Genome Atlas Project Raw Sequencing Reads: With skyrocketing numbers of whole genome sequence and phenotype data available from individuals' germline and diseased cells, we need a new framework for understanding genomics data. Using the Data-Scope (a data-intensive supercomputer, funded by the NSF), we aim to detect sets of nucleotide-level variations that best classify given phenotypes. Next, we can find covarying or spatially correlated genomic variations across the entire dataset or within phenotypes. Our final goal, and the most powerful application of these data and algorithms, is to use unsupervised methods to delineate genomic variants that discriminate subsets of the data, without regard to phenotypes.

SEED FUNDING AWARDEES

TAMER A. ZAKI

The Elusive Onset of Turbulence And The Laminar-Turbulence Interface: The onset of chaotic fluid motion from an initially laminar, organized state is an intriguing phenomenon referred to as laminar-to-turbulence transition. Early stages involve the amplification of seemingly innocuous small-amplitude perturbations. Once these disturbances reach appreciable amplitudes, they become host to sporadic bursts of turbulence — a chaotic state whose complexity is only tractable by high-fidelity large-scale simulations. By performing direct numerical simulations that resolve the dynamics of laminar-to-turbulence transition in space and time, and storing the full history of the flow evolution, we capture the rare high-amplitude events that give way to turbulence and unravel key characteristics of the laminar-turbulence interface.

BEN LANGMEAD

Highly Scalable Software for Analyzing Large Collections of RNA Sequencing Data: We are developing a radically scalable software tool, Rail-RNA, for analysis of large RNA sequencing datasets. Rail-RNA will make it easy for researchers to re-analyze published RNA-seq datasets. It will be designed to analyze many datasets at once, applying an identical analysis method to each so that results are comparable. This enables researchers to perform several critical scientific tasks that are currently difficult, including (a) reproducing results from previous large RNA-seq studies, (b) comparing datasets while avoiding bioinformatic variability, (c) studying systematic biases and other effects (e.g lab and batch effects) that can confound conclusions when disparate datasets are combined.

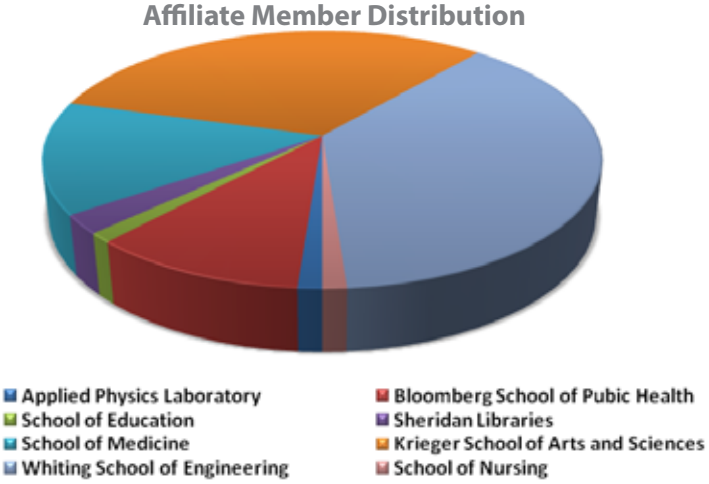
NITIN DAPHALAPURKAR

Professors Daphalapurkar and Graham-Brady of Hopkins Extreme Materials Institute are constructing a massive dynamic-fragmentation database (**FragData**) for materials undergoing failure in critical applications. They envisage FragData would help expand understanding on the mechanics of failure processes associated with, for example, disruption of asteroids, fragmentation of protection materials under impact, and debris formation of construction materials under catastrophic loading. The idea is to have the database openly accessible, have tools to carry out in situ analysis, and have the database serve as a central platform for other researchers to interpret the massive data from state-of-the-art particle-based and finite-element-based simulation techniques.

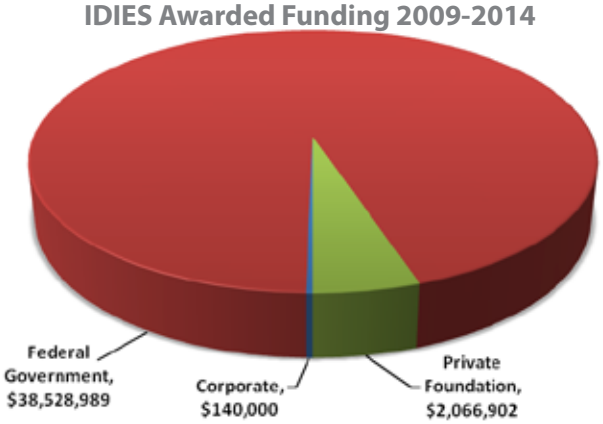
IDIES IN NUMBERS

IDIES affiliate members belong to eight of JHU's schools and divisions, supporting IDIES's Vision: to facilitate high performance and data intensive computing across all of JHU. Since its inception in 2009, IDIES has grown to include affiliate members with diverse research interests. IDIES research proposal submissions have experienced an overall positive trend averaging an increase of 50% each year.

Affiliate Member Distribution



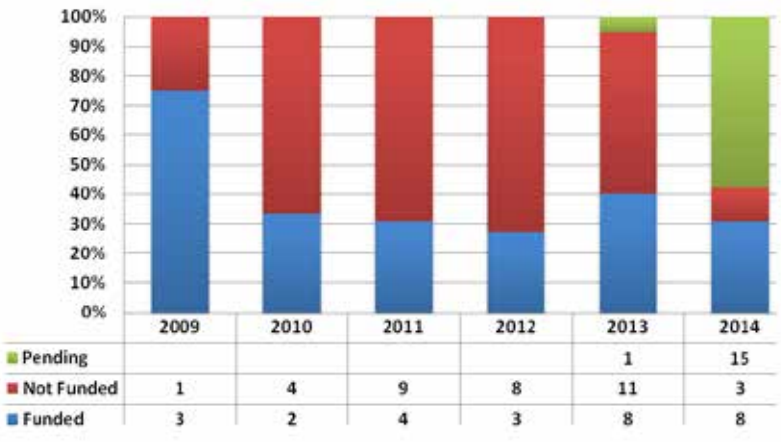
IDIES Awarded Funding 2009-2014



Currently, 79% of awarded IDIES proposals are supported by programs within the federal government. The important work being completed by IDIES also attracts the interest of private foundations (14%) and corporate sponsors (7%). Affiliate proposal submissions through IDIES have resulted in collective sponsored awards totalling $40,735,891 for the period 2009 through 2014.

IDIES research proposals averaged a success rate of 44% from FY 2009 through 2014, which compares favorably with FY 2013 award rates of 22% and 14.3% reported by the NSF[1] and NIH[2], respectively. As the research environment continues to experience a restriction in available funding, sponsoring agencies are targeting research projects that present a multidisciplinary approach led by a diversified team of leading experts[3]. IDIES has created an ideal situation for these teams to form at JHU, bringing together research groups from across the institution through the common study of big data science.

IDIES Proposal Success Rate



	2009	2010	2011	2012	2013	2014
Pending					1	15
Not Funded	1	4	9	8	11	3
Funded	3	2	4	3	8	8

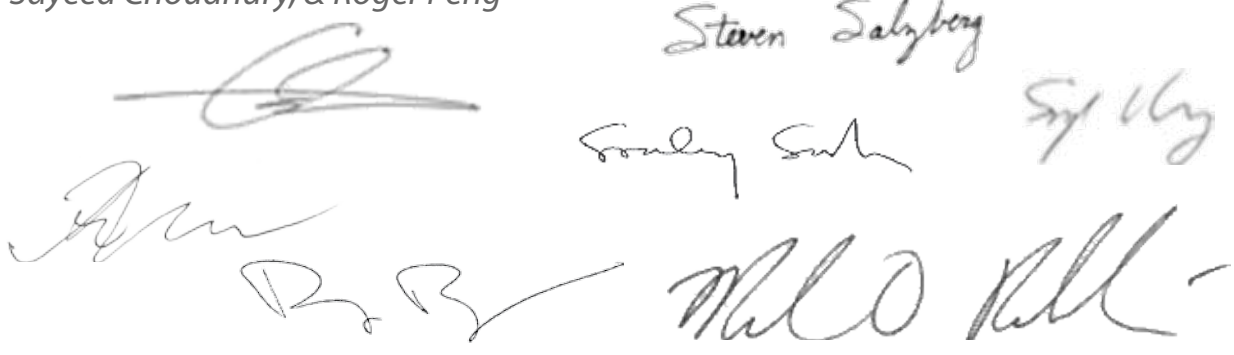
1. FY 2013, Funding by State and Organization, NSF.gov
2. FY13 R01 success rate, Research Project Grants (RPG) and Other Mechanisms, report, NIH.gov
3. NCURA NIH Update, 2014

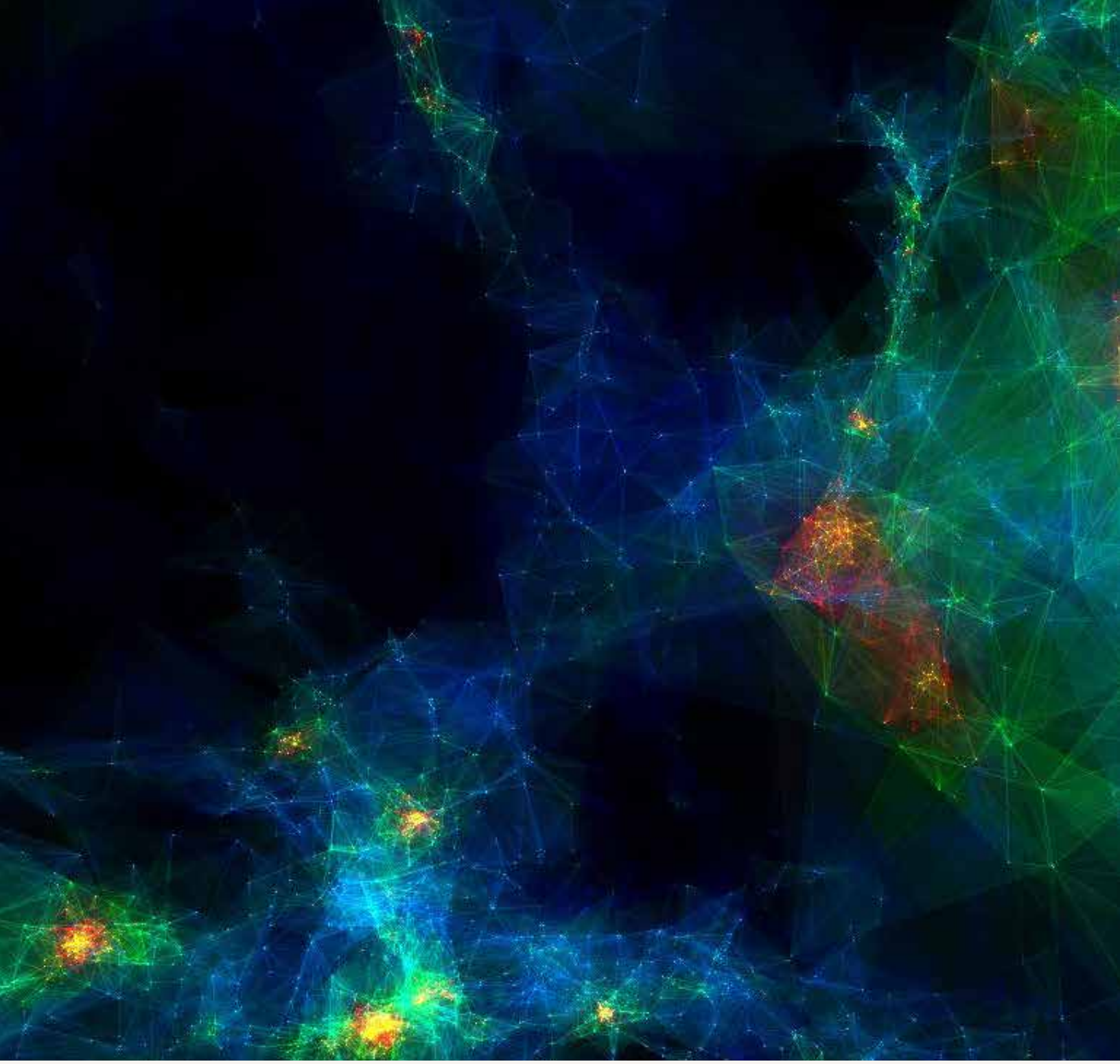
THANK YOU

TO OUR GENEROUS SPONSORS

The IDIES Executive Committee would like to extend our heartfelt gratitude to our affiliates, collaborators, contributors, editors, and staff, without whose continued support and cooperation IDIES would not be possible.

—Alex Szalay, Charles Meneveau, Stephen Salzberg, Mark Robbins , Ani Thakar, Sayeed Choudhury, & Roger Peng





JOHNS HOPKINS
INSTITUTE FOR
DATA-INTENSIVE
ENGINEERING & SCIENCE

IDIES • Johns Hopkins University • 3400 N. Charles St • Baltimore, MD 21218